Maier, Uwe; Randler, Christoph; Wolf, Nicole

Effekte von computergestützten, formativen Tests mit unterschiedlichen Rückmeldeformaten auf Lernleistungen im naturwissenschaftlichen Unterricht

Zeitschrift für Pädagogik 62 (2016) 2, S. 241-262

D 7484

ZEITSCHRIFT FÜR PÄDAGOGIK

Heft 2 März/April 2016

BELTZ JUVENTA

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 62 – Heft 2
März/April 2016

Inhaltsverzeichnis

Thementeil: Konzeptualisierungen des Biografischen – Zur Aktualität biografiewissenschaftlicher Perspektiven in der Pädagogik

Allgemeiner Teil

Essay

Besprechung

Dokumentation

Table of Contents

Uwe Maier/Christoph Randler/Nicole Wolf

Effekte von computergestützten, formativen Tests mit unterschiedlichen Rückmeldeformaten auf Lernleistungen im naturwissenschaftlichen Unterricht

Zusammenfassung: Formative Leistungsmessungen werden in der Lehr-Lernforschung als effektive Methode zur Steigerung von Schülerleistungen beschrieben. Allerdings variieren die Effekte stark in Abhängigkeit von Lerninhalt, Diagnoseverfahren und Art der Rückmeldung. In einer randomisierten, experimentellen Studie im Fach Biologie wurde deshalb der Frage nachgegangen, ob bei einem computergestützten, formativen Leistungstest ausführliche Rückmeldungen (Treatment 1) zu besseren Lernergebnissen führen als einfache Rückmeldungen (Treatment 2). In einer Kontrollgruppe lasen die Schülerinnen und Schüler inhaltsgleiche Texte. An der Studie nahmen zehn Schulklassen (Sek I) mit insgesamt 261 Schülerinnen und Schülern teil. Es zeigte sich, dass einfache Rückmeldungen sowohl im Posttest als auch im Behaltenstest besser abschneiden als das Treatment 1 bzw. die Kontrollgruppe. Berücksichtigt man allerdings die Nutzung der Rückmeldungen, so lassen sich positive Leistungseffekte der ausführlichen Rückmeldung im Vergleich zur Kontrollgruppe nachweisen.

Schlagworte: formative Leistungsdiagnostik, Tablets, Rückmeldung, naturwissenschaftlicher Unterricht, Schülerleistung

1. Theoretischer Hintergrund

1.1 Formative Leistungsmessung im naturwissenschaftlichen Unterricht

Unter formativer Leistungsmessung versteht man die Einbettung standardisierter oder informeller Verfahren der Leistungsdiagnostik in den täglichen Unterrichtsverlauf mit dem Ziel der Adaption von Unterricht. Komponenten einer formativen Leistungsdiagnostik sind (Black & Wiliam, 2009; Maier, 2010; Clark, 2012; Maier, Hofmann & Zeitler, 2012):

a) Die Lehrkraft definiert ein Lernziel und konkrete Indikatoren, die Auskunft über das Erreichen des Lernziels geben.
b) Es werden in den Unterricht integrierte Leistungssituationen zur Überprüfung der Lernzielerreichung genutzt.
c) Die Lernenden erhalten eine Rückmeldung, die sich auf die Lernziele und die Erfolgsindikatoren bezieht.
d) Die Schülerinnen und Schüler sollen zunehmend in die Lage versetzt werden, ihre Lernfortschritte selbst erfassen und bewerten zu können.

e) Der Unterricht wird adaptiv gestaltet und orientiert sich an den formativ diagnostizierten Lernvoraussetzungen der Schülerinnen und Schüler. Der Erfolg von Fördermaßnahmen kann durch wiederholte formative Leistungsdiagnosen evaluiert werden.

Studien und Metaanalysen aus der empirischen Unterrichtsforschung zeigen, dass Diagnoseverfahren und Förderansätze, die sich an diesen Komponenten formativer Leistungsdiagnostik orientieren, mittlere bis hohe Effekte auf Lernzuwächse haben können (Fuchs & Fuchs, 1986; Crooks, 1988; Fraser, Walberg, Welch & Hattie, 1987; Black & Wiliam, 1998; Hattie, 2009; Kingston & Nash, 2011). Die Effekte können allerdings von Verfahren zu Verfahren und von Fach zu Fach erheblich schwanken. Kingston und Nash (2011) fanden in ihrer Metaanalyse heraus, dass formative Diagnostik in sprachlichen Fächern zu wesentlich höheren Effektstärken führt als in Mathematik oder den naturwissenschaftlichen Fächern. Die Autoren erklären dieses Ergebnis mit der Feedbackinterventionstheorie (Kluger & DeNisi, 1996), die davon ausgeht, dass Rückmeldungen bei eher bekannten und einfachen Aufgaben (wie z. B. Vokabeln lernen oder Grammatik) effektiver sind als bei komplexeren Aufgaben. Bennett (2011) geht noch einen Schritt weiter und kritisiert, dass sich Studien und Metaanalysen bisher zu wenig an domänenspezifischen Eigenheiten formativer Diagnostik orientieren.

Die hier vorliegende Studie wurde im Biologieunterricht durchgeführt und orientiert sich deshalb an Verfahren der formativen Leistungsmessung in den naturwissenschaftlichen Fächern. In diesen spielt der Erwerb von konzeptuellem Begriffswissen eine wichtige Rolle (Duit, 2003; Posner, Strike, Hewson & Gertzog, 1982; Duit & Treagust, 2010). Lerntheoretisch stellt man sich den Erwerb von naturwissenschaftlichen Konzepten als aktiven Konstruktionsvorgang der Schülerinnen und Schüler vor. Dabei sind vorwissenschaftliche Begriffe eine wichtige Grundlage für den Konzeptwechsel. In der naturwissenschaftsdidaktischen Literatur finden sich zahlreiche Vorschläge für Verfahren der Diagnostik von konzeptuellem Wissen (Ruiz-Primo & Shavelson, 1996; Chang, Sung, Chang & Lin, 2005; Anderson, Zuiker, Taasoobshirazi & Hickey, 2007; McConnell et al., 2006). Diese Verfahren können im Rahmen einer formativen Leistungsmessung genutzt werden, z. B. zur Erfassung des Vorwissens oder zur Überprüfung von Lernfortschritten (Yin et al., 2008; Furtak, 2012). Concept-Maps oder Essay-Aufgaben eröffnen den Lehrkräften einen besonders genauen Einblick in die Vorstellungswelt der Schülerinnen und Schüler. Von Nachteil ist allerdings der hohe Durchführungs- und Auswertungsaufwand, der einen breiten Einsatz im naturwissenschaftlichen Unterricht bisher eher verhindert. Besonders zeitökonomisch sind dagegen Tests mit geschlossenen Fragen zur Erfassung des Begriffswissens (z. B. Donovan, 2008). Nachteile dieser Tests sind die hohe Ratewahrscheinlichkeit sowie die Schwierigkeit, anspruchsvolle und valide Testfragen zur Überprüfung des konzeptuellen Wissens zu konstruieren.

Ein Versuch, diese Problematik zumindest ansatzweise zu lösen, sind ‚two-tier diagnostic assessments' (Treagust, 1988; Lin, 2004; Chandrasegaran, Treagust & Mocerino, 2007). Diese Tests bestehen aus einer Reihe von zweischrittigen Items. Im Inhaltsteil wird nach einem naturwissenschaftlichen Phänomen oder nach dem Ausgang eines na-

turwissenschaftlichen Experiments gefragt (z. B.: Was passiert mit einem Lichtstrahl beim Übergang von Luft in Wasser?). Die Schülerinnen und Schüler erhalten verschiedene Alternativen für den Ausgang des Experiments zur Auswahl. Im zweiten Teil werden sich widersprechende Begründungen zu den vorangehenden Alternativen offeriert. Wenn ein Proband beide Teile des Items richtig löst, kann man mit großer Wahrscheinlichkeit davon ausgehen, dass nicht nur Fakten auswendig gelernt wurden (oder zufällig geraten wurde), sondern dass vertieftes Begriffswissen vorhanden ist.

Es gibt eine Reihe von Studien, die eine computergestützte Umsetzung von ‚two-tier diagnostic assessments' erprobten und gleichzeitig evaluierten, wie sich deren Einsatz auf die Lernleistung im Unterricht auswirkt. Ein Beispiel ist das Projekt DIAGNOSER, ein webbasiertes Diagnosetool zum Aufspüren von Schülerfehlvorstellungen bei physikalischen Konzepten in der Sekundarstufe (Thissen-Roe, Hunt & Minstrell, 2004). Die Testfragen wurden von Physiklehrkräften im Bundesstaat Washington entwickelt und von den Schülerinnen und Schülern über eine Online-Plattform bearbeitet. Nach der Testdurchführung wurden Übungsaufgaben angeboten. In den jährlichen, zentralen Leistungstests hatten Schulen, die mit dem Testsystem arbeiteten, 14 % bessere Leistungsergebnisse im Vergleich zum Landeswert. Allerdings handelte es sich dabei um eine selbstselektive Gelegenheitsstichprobe, weil der Einsatz von DIAGNOSER freiwillig war. Auch Lai und Chen (2010) fanden in einer quasiexperimentellen Studie in einer Primarschule in Taiwan einen positiven Lerneffekt beim Einsatz eines Two-tier-Tests zu Begriffen und Phänomenen der Elektrizitätslehre. Auch hier wurde die Testrückmeldung mit passenden Übungsaufgaben gekoppelt.

1.2 Feedback als zentrales Element formativer Leistungsmessung

Feedback ist die Schnittstelle zwischen Leistungsmessung und weiterführendem Unterricht bzw. individuellen Förderaktivitäten (Lysakowski & Walberg, 1982; Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Mory, 1992; Kluger & DeNisi, 1996; Hattie & Timperley, 2007). Feedback an die Lehrkräfte kann idealerweise dazu führen, dass der Erfolg des bisherigen Unterrichts eine sichtbare Bestätigung findet bzw. im weiteren Unterrichtsverlauf auf Wissenslücken reagiert wird. Feedback an die Schülerinnen und Schüler kann dazu führen, dass entweder erworbenes Wissen bestätigt wird oder konzeptuelle Fehlvorstellungen erkannt werden und selbständig nach anderen Lösungsmöglichkeiten gesucht wird. Im Sinne des selbstgesteuerten Lernens wäre es ideal, wenn Lernende nach und nach in die Lage versetzt werden, die Rückmeldungen aus formativen Leistungsmessungen eigenständig zu interpretieren und für den weiteren Lernverlauf zu nutzen (Clark, 2012). Bangert-Drowns et al. (1991) weisen allerdings darauf hin, dass Lernen auch ohne Feedback stattfinden kann, beispielsweise beim Beobachtungslernen. Rückmeldungen können zudem lernhinderlich sein, beispielsweise wenn die Aufgaben zu einfach sind und Rückmeldeinformationen eher ablenken (Mory, 1992). Aus diesen Gründen ist es relevant, die Art der Rückmeldung genauer zu beschreiben.

Es gibt eine sehr umfangreiche Forschung zu Feedbackeffekten. Generell lässt sich auf Basis von Metaanalysen und Literaturübersichten sagen, dass Rückmeldungen im Rahmen von Lehr-Lernprozessen lernförderlich sein können (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996; Hattie & Timperley, 2007; Hattie, 2009). Es gibt jedoch eine Reihe von Mediatorvariablen, die die Höhe des Feedbackeffekts beeinflussen. Studien zur Bedeutung dieser Mediatorvariablen werden in den folgenden Absätzen skizziert.

Kluger und DeNisi (1996) weisen darauf hin, dass Feedback in komplexen Lerndomänen bzw. bei komplexen Aufgabentypen zu geringeren Effekten führt als bei einfachen Aufgabentypen. Erklärt wird dies damit, dass es bei einfachen Aufgaben (z. B. Abfrage von Faktenwissen) dem Lernenden eher gelingt, die Feedbackinformation mit seiner Aufgabenleistung zu verknüpfen. Bangert-Drowns et al. (1991) erklären unterschiedliche Effektstärken in ihrer Metaanalyse unter anderem mit dem Informationsgehalt. Es finden sich keine Effekte bei einfachem Richtig/Falsch-Feedback, jedoch höhere Effekte, wenn die richtige Antwort rückgemeldet wird. McKendree (1990) fand Vorteile eines ausführlichen Feedbacks bei der Erprobung eines computerbasierten Geometrie-Tutors. Die Lernleistung erhöhte sich, wenn den Schülerinnen und Schülern bei Fehlern mitgeteilt wurde, gegen welche Regeln verstoßen wurde und worauf beim nächsten Lösungsversuch zu achten war. Nagata (1993) zeigte mit einem tutoriellen System zum Sprachenlernen, dass elaboriertes Feedback zu höheren Lerneffekten führt als die einfache Rückmeldung der Korrektheit. Ebenso gibt es Hinweise, dass ein detailliertes Feedback zu Schreibaufgaben zu besseren Lernleistungen bei weiteren Schreibaufgaben führen kann (Lipnevich & Smith, 2009). Bürgermeister et al. (2011) zeigen, dass lösungsprozessbezogenes Feedback im Vergleich zum sozial vergleichenden Feedback sowohl die Motivation als auch die Entwicklung der mathematischen Modellierungskompetenz fördern kann.

Feedback im Rahmen computergestützter bzw. programmierter Instruktion wird als eher gering eingeschätzt. Ein Problem der Metaanalysen ist allerdings, dass sie ausschließlich auf ältere Studien zurückgreifen. Bangert-Drowns et al. (1991) stützen ihre Analysen auf Studien zu Feedback im Rahmen der programmierten Instruktion der 1960er- und 1970er-Jahre. Auch die Befunde zu Feedback im Rahmen eines computergestützten Unterrichts aus dieser Zeit können nicht unbedingt auf heute übertragen werden. Ein weiteres Problem der Feedbackforschung ist, dass viele Befunde auf Laborstudien zurückgehen, was die externe Validität verringert.

Aktuelle Studien zu Feedback deuten auf die Bedeutung der Nutzung des Feedbacks durch die Lernenden hin. Timmers und Veldkamp (2011) zeigten, dass die Nutzung von Rückmeldungen in einem computergestützten Setting sehr stark individuell variiert und Lernende vor allem den nicht richtig gelösten Aufgaben Beachtung schenken. In einer experimentellen Unterrichtsstudie zu Feedback in Mathematik fand sich ein direkter Effekt von lösungsprozessorientiertem Feedback im Vergleich zu sozial vergleichendem Feedback auf Wissenserwerb und Motivation (Rakoczy, Klieme, Bürgermeister & Harks, 2008; Rakoczy, Harks, Klieme, Blum & Hochweber, 2013). Es gab allerdings einen indirekten Effekt von informativem Feedback auf die Schülerleistung über die von den Schülerinnen und Schülern wahrgenommene Nützlichkeit des Feedbacks.

1.3 Möglichkeiten der computergestützten, formativen Leistungsmessung

Sowohl das Internet als auch die zunehmende Verfügbarkeit von mobilen Endgeräten haben der Entwicklung anspruchsvoller und zugleich praktikabler Verfahren der formativen Diagnostik Vorschub geleistet (Russell, 2010; Maier, 2014). Eine Möglichkeit für die schulische Realisierung formativer Diagnostik sind Lernplattformen wie Moodle. Diese bietet mit der Funktionalität ‚Test' eine Möglichkeit zur Konstruktion und Durchführung formativer Leistungsmessungen im Unterricht. Lehrkräfte können unterschiedliche Testaufgabenformate wählen und mit der Zeit eine umfangreiche Aufgabensammlung anlegen. Die Prüfungsaufgaben lassen sich zu einzelnen Tests kombinieren und wiederholt durchführen. Die Ergebnisse werden auf Individual- und Klassenebene übersichtlich dargestellt. Darüber hinaus lassen sich unterschiedliche Formen von Feedback für die Lernenden programmieren. Lehrkräfte können auf der Lernplattform Übungs- und Wiederholungsmaterialien zur Verfügung stellen.

Die empirische Befundlage zur Lernwirksamkeit von Feedback über Lernplattformen ist noch sehr dünn. Es gibt vereinzelt Hinweise auf positive Effekte, wie z. B. im Bereich des Lernens von Vokabeln im Englischunterricht (Jia, Chen, Ding & Ruan, 2012). Ebenso konnte gezeigt werden, dass eine Reihe von Feedbackstrategien das selbstregulierte Lernen unterstützen können (z. B. Wang, 2011). Die Entwicklung und zunehmende Verbreitung von mobilen Endgeräten eröffnet weitere Perspektiven für eine Umsetzung computergestützter, formativer Leistungsdiagnostik an Schulen. Die Bildschirmdarstellung von Moodle lässt sich für mobile Endgeräte optimieren, sodass einfachere Testaufgaben auch auf Smartphones oder Tablets bearbeitet werden können. Auch wenn bisher nur einzelne Schulen ihre Klassen mit Tablets ausstatten, wird sich dieser Trend in Zukunft verstärken.

2. Hypothesen

Ziel dieser Studie ist die Entwicklung und Erprobung einer computergestützten, formativen Leistungsmessung für den naturwissenschaftlichen Unterricht mit Tablets. Dabei soll eine einzelne Komponente formativer Diagnostik gezielt erprobt und variiert werden. Zur Erfassung des Wissens wird auf ein bewährtes Instrument in den Naturwissenschaftsdidaktiken zurückgegriffen, auf mehrschrittige, geschlossene Testaufgaben mit Inhalts- und Begründungsteilen (two-tier diagnostic tests). Dabei wird der Frage nachgegangen, welchen Effekt verschiedene Rückmeldeformen zu den Testantworten auf die Lernleistung haben. Die bisherigen Befunde der Feedbackforschung sind hierzu nicht eindeutig. Tendenziell gibt es jedoch Hinweise, dass vor allem bei Aufgabenstellungen zu konzeptuellem Wissen eine ausführliche Rückmeldung zur Lösung effektiver ist als die einfache Rückmeldung zur Korrektheit der Antwort. Diese Annahme basiert jedoch vorwiegend auf älteren Laborstudien in unterschiedlichen Lernkontexten. Ebenso gibt es bisher noch keine Studie, in der Leistungseffekte von Rückmeldungen im Rahmen von computergestützten ‚two-tier diagnostic tests' erkundet wurden.

Es sollen folgende Hypothesen überprüft werden:

1) Schülerinnen und Schüler, die während einer Unterrichtseinheit in einem naturwissenschaftlichen Fach formative Leistungstests bearbeiten und dabei eine ausführliche, schriftliche Rückmeldung nach jedem Item erhalten, erzielen bei einem abschließenden Leistungstest (zur Erfassung von Faktenwissen und konzeptuellem Wissen) einen höheren Punktwert als Schülerinnen und Schüler, die während der gleichen Unterrichtseinheit die gleiche Anzahl formativer Leistungstests (FLT) bearbeiten, jedoch lediglich am Ende des Tests angezeigt bekommen, welche Aufgaben korrekt bzw. nicht korrekt gelöst wurden.
2) Schülerinnen und Schüler, die während einer Unterrichtseinheit in einem naturwissenschaftlichen Fach formative Leistungstests (FLT) mit einfachen oder ausführlichen Rückmeldungen bearbeiten, erzielen bei einem abschließenden Leistungstest (zur Erfassung von Faktenwissen und konzeptuellem Wissen) einen höheren Punktwert als Schülerinnen und Schüler, die keine formativen Tests bearbeiten und dafür Texte zum Thema lesen.

3. Methodisches Vorgehen

3.1 Unterrichtseinheit

Als Lerninhalt für die Überprüfung der Fragestellung wurde die Unterrichtseinheit ‚Anpassung der Vögel an den Lebensraum Luft' ausgewählt. Für diese Unterrichtseinheit lag ein fertiges Planungskonzept vor, das in empirischen Studien erprobt wurde (Randler & Hummel, 2011). Für die teilnehmenden Lehrkräfte wurde eine Unterrichtsplanung erstellt, die sich auf sechs Wochen mit je zwei Biologiestunden pro Woche erstreckte. Der Unterricht wurde in folgende Themenblöcke gegliedert: (1) Gemeinsame Anpassungsmerkmale von Vögeln und der Begriff der evolutionären Anpassung; (2) Anpassung der Vögel an das Fliegen, Flugformen und das Konzept des Auftriebs; (3) Vogelzug. Stundenskizzen, sämtliche Arbeitsblätter sowie Materialien für einen Lernzirkel mit Experimenten zum Vogelflug wurden allen Lehrkräften zur Verfügung gestellt.

Ziel der Unterrichtseinheit war einerseits der Erwerb von Faktenwissen zur Anpassung von Vögeln (z. B. unterschiedliche Schnabel- und Fußformen, Arten des Vogelflugs etc.) und andererseits der Erwerb von konzeptuellem Wissen (Begriffe wie ‚evolutionäre Anpassung' oder ‚Auftrieb'). Faktenwissen und konzeptuelles Wissen können dem deklarativen Wissen zugeordnet werden. Deklaratives Wissen besteht aus propositionalen Netzwerken, kann als ‚Weltwissen' bezeichnet werden und ist im Vergleich zum prozeduralen Wissen (Routinen, Automatismen) verbalisierbar (Steiner, 2001). Für eine unterrichtspraktische Unterscheidung von Faktenwissen und konzeptuellem Wissen bietet sich die revidierte Bloom'sche Lernzieltaxonomie von Anderson und Krathwohl (2001) an. Faktenwissen wird als singuläre Verknüpfungen einzelner Wissenselemente

beschrieben, während sich konzeptuelles Wissen durch eine komplexe Vernetzung mit verschiedenen Wissenselementen auszeichnet (sog. Begriffsnetze).

3.2 Leistungstests

Um das Vorwissen der Schülerinnen und Schüler bzw. deren Lernverlauf erfassen zu können, wurden zu den einzelnen Themenbereichen der Unterrichtseinheit formative Leistungstests entwickelt. Technisch wurden die benötigten Prüfungsaufgaben mit der Funktionalität ‚Test' der Lernplattform Moodle (Version 2.0) umgesetzt. Um den Test für die Schülerinnen und Schüler abwechslungsreich zu gestalten, wurden verschiedene Aufgabenformate gewählt: Ja/Nein-Fragen, Single-Choice-Aufgaben, Multiple-Choice-Aufgaben und Lückentextaufgaben. Vor allem mit den Lückentextaufgaben wurden die relevanten Begriffe ‚evolutionäre Anpassung' und ‚Auftrieb' abgeprüft. Dieses Aufgabenformat eignet sich, um das in der naturwissenschaftsdidaktischen Literatur beschriebene Verfahren der ‚two-tier diagnostic tests' umzusetzen.

Um die Testfragen zu konstruieren, wurde auf Studien zurückgegriffen, die konzeptuelle Schülerfehlvorstellungen zum Begriff der evolutionären Anpassung beschrieben (z. B. Halldén, 1988; Baalmann, Frerichs, Weitzel, Gropengießer & Kattmann, 2004; Nehm & Reilly, 2007; Zabel & Gropengießer, 2010). Single-Choice-Testfragen zum konzeptuellen Wissen wurden so aufgebaut, dass als Distraktoren mögliche Schülerfehlvorstellungen angeboten wurden (z. B. Anpassung ist ein intentionaler Vorgang; die stärksten Tiere setzen sich durch). Damit ist erstens gewährleistet, dass die Schülerinnen und Schüler nicht einfach durch Ausschluss von unlogisch klingenden Antwortvorgaben auf die richtige Antwort schließen können. Zweitens können fehlerhafte Antworten der Lehrkraft Hinweise auf mögliche Fehlvorstellungen geben, sodass im Sinne einer formativen Diagnostik darauf im weiterführenden Unterricht reagiert werden kann.

Mit dem Testaufgabenformat ‚Lückentext' konnten zudem zwei- bzw. mehrschrittige Testitems konstruiert werden, in denen sich die einzelnen Teilantworten aufeinander bezogen. Das Aufgabenformat ermöglicht somit die Hintereinanderschaltung mehrerer Single-Choice-Testaufgaben sowie die Verknüpfung mit Text und Bildern. Damit kann den Schülerinnen und Schülern ein konkreter Fall evolutionärer Anpassung präsentiert werden. Im weiteren Verlauf der Aufgabe kann man die Begründungen abfragen. Tabelle 1 zeigt ein Beispielitem zur Anpassung von Schmetterlingen. Im ersten Teil des Items wird zunächst ein Phänomen der evolutionären Anpassung beschrieben (wenige Schmetterlinge in einer Population sind dunkel) und gleich im Anschluss nach dem Grund für dieses Phänomen gefragt. Im zweiten Schritt sollen sich die Schülerinnen und Schüler entscheiden, was im Laufe der Zeit mit der Population passieren könnte (Inhaltsfrage). Im dritten Teil wird dann wiederum nach einer Begründung hierfür gefragt. Das Item ist so aufgebaut, dass pro Fragenteil jeweils verschiedene Schülerfehlvorstellungen als Distraktoren angeboten werden. Damit sind bei einer falschen Beantwortung zumindest ansatzweise Rückschlüsse auf Missverständnisse möglich.

Fragentext	**Antwortalternativen *(korrekte Antwort kursiv)***
In einer Population von Schmetterlingen sind fast alle Schmetterlinge hell, nur ein Schmetterling hat eine dunkle Farbe. Warum ist er dunkel?	Innerhalb einer Population kann es gar keine Schmetterlinge mit unterschiedlichen Farben geben. *Aufgrund einer zufälligen Veränderung im Erbmaterial.* Im neuen Lebensraum mit dunklen Bäumen hat sich der schlauste Schmetterling getarnt, um zu überleben.
Wenn das dunkle Merkmal die Chance zu überleben erhöhen würde, was würde dann mit der Population nach einer längeren Zeitspanne wohl passieren?	Alle Schmetterlinge würden dunkel werden, *Die Anzahl der dunklen Schmetterlinge würde sich erhöhen, es würde aber dennoch hellere Schmetterlinge geben,* Es würde sich nichts ändern,
weil	sich alle Lebewesen anpassen, um leichter zu überleben. *vorwiegend die Lebewesen mit dem dunklen Merkmal überleben und sich dann vermehren.* sich die geringe Anzahl dunkler Schmetterlinge nicht durchsetzen könnte.

Tab. 1: Beispielitem

Es wurde eine Fragensammlung angelegt, die alle Inhalte der Unterrichtseinheit abdeckte. Ein besonderer Schwerpunkt lag bei der Prüfung des Konzepts der evolutionären Anpassung. Mit der Fragensammlung wurden folgende Tests zusammengestellt:

- ein Pretest zur Erfassung des Vorwissens mit insgesamt 21 Testaufgaben (Cronbach's Alpha = .48)
- ein formativer Test zum Themenblock ‚Anpassungsmerkmale von Vögeln und Anpassungsbegriff' mit insgesamt 16 Testaufgaben (FLT 1, Cronbach's Alpha = .80)
- ein formativer Test zum Themenblock ‚Anpassung an das Fliegen, Vogelflug und Auftrieb' mit insgesamt 14 Testaufgaben (FLT 2, Cronbach's Alpha = .77)
- ein mit dem Pretest identischer Posttest zur Erfassung des Wissenszuwachses während der Unterrichtseinheit mit insgesamt 21 Testaufgaben (Cronbach's Alpha = .80)

Die internen Konsistenzen der Leistungstests sind bis auf den Pretest gut. Der niedrige Alpha-Wert des Pretests hängt vor allem mit dem geringen Vorwissen der Schülerinnen und Schüler zusammen (Randler, 2012). Dies ist dadurch bedingt, dass die Schülerinnen und Schüler wenig Vorwissen haben und deswegen die Fragen größtenteils falsch beantworten. Allerdings sind die falschen Antworten nicht konsistent falsch, sondern es werden bei Disktraktoren mal die einen, mal die anderen Antworten angekreuzt. Durch den Einsatz eines Tests zur längsschnittlichen Erfassung des Wissenszuwachses sind Störeffekte durch Testwiederholungen nicht auszuschließen (Bortz & Döring, 2006; Hussy, Schreier & Echterhoff, 2010). Vor allem Übungs- und Erinnerungseffekte könnten sich positiv auf die abhängige Variable auswirken. Aufgrund der langen Zeitspanne zwischen Pre- und Posttest (ca. sechs Wochen) sind Erinnerungseffekte jedoch eher un-

wahrscheinlich. Zudem wurden die Lösungen des Pretests mit den Schülerinnen und Schülern nicht besprochen. Sollten dennoch Störeffekte aufgrund der Testwiederholung auftreten, müssten sich diese aufgrund der Randomisierung gleichermaßen auf alle drei Versuchsgruppen auswirken.

Zur Überprüfung des langfristigen Wissenserwerbs und auch zur Validierung der elektronischen Tests wurde ein Behaltenstest mit halboffenen Fragen im Paper-Pencil-Format entwickelt. Inhaltlich deckt dieser den Pre- bzw. Posttest ab. Es wurde darauf geachtet, dass die Items strukturell ähnlich aufgebaut sind. Die Schülerinnen und Schüler müssen allerdings ihr Wissen in Form kurzer Antwortsätze dokumentieren. In einem Beispielitem geht es um die Entwicklung der Halslänge von Giraffen. Ein Bild zeigt Giraffen mit verschiedener Halslänge und Bäume mit Blättern auf einer bestimmten Höhe. Die Schülerinnen und Schüler sollen beschreiben, wie sich die Halslänge der Giraffen im Laufe der Zeit entwickeln wird, und dies schriftlich begründen.

Der Behaltenstest wurde von 246 Schülerinnen und Schülern sechs bis acht Wochen nach dem Posttest geschrieben. Um eine möglichst einheitliche Auswertung zu gewährleisten, wurden zuvor im Team ein einheitlicher Erwartungshorizont inklusive der Kodierung erarbeitet und konkrete Lösungen für die einzelnen Testaufgaben formuliert. Die Korrektur der Arbeiten erfolgte durch eine Person. Zu Beginn der Korrektur wurden im Team Zweitkorrekturen angefertigt, um die Bewertung der Lösungen weiter zu präzisieren. Am Ende wurden die Korrekturen stichprobenartig kontrolliert. Posttest und Behaltenstest korrelieren sehr hoch ($r = .68$; $p < .001$), was für eine hohe Validität des computergestützten Posttests spricht.

3.3 Variation der Rückmeldungen

Moodle bietet zudem die Möglichkeit, unterschiedliche Rückmeldungen für einzelne Aufgabenformate zu programmieren. Diese Funktionalität wurde genutzt, um die Menge der mit dem Feedback transportierten Informationen zu variieren. Beim Pre- und Posttest erhielten die Schülerinnen und Schüler keine aufgabenbezogenen Rückmeldungen. Sie konnten lediglich am Ende des Tests sehen, wie viel Prozent der Antworten korrekt waren. Variiert wurden dagegen die aufgabenbezogenen Rückmeldungen bei den formativen Leistungstests 1 und 2. Es wurden jeweils zwei Versionen in Moodle eingepflegt. Bei einer ersten Version (Treatment 1) erhielten die Schülerinnen und Schüler nach jeder falsch gelösten Aufgabe eine ausführliche Rückmeldung. Auf dem Display wurde oben der von den Schülerinnen und Schülern ausgefüllte Lückentext weiterhin dargestellt. Falsche Teilantworten wurden rot markiert, richtige grün. Zusätzlich wurde unter dem Item in einem gesonderten Kasten ein ausführlicher Feedbacktext eingeblendet, zum Beispiel wurde für das oben dargestellte Beispiel folgende Rückmeldung eingeblendet:

> „Eine zufällige Veränderung im Erbmaterial sorgt zunächst dafür, dass bei einzelnen Schmetterlingen die Flügel dunkel sind. Da dieses Merkmal vorteilhaft ist (Tar-

nung), werden vorwiegend Schmetterlinge mit dunklen Flügeln überleben und sich vermehren. Die Anzahl der dunklen Schmetterlinge steigt. Dennoch würde es weiterhin hellere Schmetterlinge geben, denn durch Variation im Erbmaterial wird nie die ganze Population gleich aussehen."

Am Ende des Tests wurde von der Lernplattform eine Übersicht zu allen Aufgaben präsentiert. Falsch gelöste Aufgaben wurden rot markiert, richtig gelöste Aufgaben grün. Teilweise richtig beantwortete Fragen wurden gelb markiert. Die Schülerinnen und Schüler konnten innerhalb dieser Übersicht noch einmal die einzelnen Aufgaben direkt anklicken und die Rückmeldungen einsehen. Ebenso wurde der Prozentsatz richtig gelöster Aufgaben angegeben.

Bei einer zweiten Version der formativen Tests (Treatment 2) wurde auf die ausführlichen Rückmeldungen verzichtet. Nach jeder Aufgabe wurde lediglich angezeigt, ob das Item korrekt gelöst wurde oder nicht bzw., bei mehrschrittigen Items, ob es teilweise korrekt gelöst wurde. Nach Abschluss des Tests wurde der Prozentsatz richtig gelöster Aufgaben angezeigt.

3.4 Fragebogeninstrumente

Um die Bildung der Versuchsgruppen zu kontrollieren, wurden intrinsische Motivation und Vornoten mithilfe eines Fragebogens direkt im Anschluss an den Pretest erfasst. Für die Erfassung der intrinsischen Motivation wurde auf Skalen zurückgegriffen, die zwischen einer Interessens- und einer Kompetenzkomponente unterscheiden (Wild, Krapp, Schiefele, Lewalter & Schreyer, 1995). Die internen Konsistenzen beider Teilskalen sind gut (Cronbach's Alpha im Vortest .85 und im Nachtest .86). Beispielitems: Im Fach Biologie lerne und beteilige ich mich am Unterricht, weil die Unterrichtsinhalte meinen Neigungen entsprechen/…weil es mir wichtig ist, meine fachlichen Fähigkeiten immer mehr zu erweitern. Als Kovariaten im Leistungsbereich wurden die letzten Schulnoten in den Fächern Deutsch, Mathematik und Naturwissenschaft erfragt.

Nach Abschluss der Unterrichtseinheit (im Rahmen des Posttests) wurden die Schülerinnen und Schüler der Treatmentgruppe 1 nach dem Nutzen und der Nutzung der ausführlichen Rückmeldungen befragt. Die Skala ‚Feedbacknutzung' bestand aus vier Items und hatte eine zufriedenstellende interne Konsistenz (Cronbach's Alpha von .75): „(1) Ich habe das Feedback ausführlich gelesen. (2) Mir war nur wichtig, ob ich richtig oder falsch geantwortet habe – die genaue Rückmeldung war mir egal (–). (3) Ich fand die weiterführenden Informationen gut. (4) Das Feedback hat mir sehr geholfen."

3.5 Stichprobe

Insgesamt beteiligten sich zehn Klassen mit 261 Schülerinnen und Schülern aus einer Mittelschule, einer Volksschule, drei Realschulen und einem Gymnasium in Nordbayern im Schuljahr 2012/13 an der Studie. Größe und Lage der Schulen waren sehr unterschiedlich. Die Volksschule lag im ländlichen Bereich und war mit ca. 220 Schülerinnen und Schülern (teilweise einzügig) eher klein. Das Gymnasium (Kleinstadt) hingegen hatte über 1500 Schülerinnen und Schüler. Die Real- und Mittelschulen befanden sich, wie das Gymnasium, in kleineren Städten. Lediglich eine Realschule, die mit drei Klassen an der Studie teilnahm, befand sich in einer Großstadt. Da an der Mittelschule eine Klasse mit dem Abschlussziel ‚Mittlere Reife' und eine Klasse mit dem Abschlussziel ‚Qualifizierter Hauptschulabschluss' teilnahmen, wurde der angestrebte Abschluss als Stichprobenmerkmal herangezogen (Abitur: 11.1 %, Mittlere Reife: 63.2 % und Qualifizierter Hauptschulabschluss: 25.7 %). Es nahmen Schülerinnen und Schüler aus den Jahrgangsstufen 6 (31.8 %) und 7 (68.2 %) teil. Der Anteil der Mädchen war mit 55.6 % leicht höher als der Jungenanteil mit 44.4 %. Ungefähr zwei Drittel der Schülerinnen und Schüler gaben an, dass bei ihnen zu Hause nur Deutsch gesprochen wird (64.1 %); ein Drittel spricht Deutsch und eine andere Sprache (32.8 %); lediglich 3.1 % gaben an, nur eine andere Sprache zu Hause zu sprechen.

Die Auswahl der Schulen bzw. Klassen kann als Gelegenheitsstichprobe bezeichnet werden. Die Autoren nutzten persönliche Kontakte oder institutionelle Kooperationen mit den Schulen. Bei der Umsetzung eines sehr stark vorstrukturierten Unterrichtsforschungsdesigns spielt die Freiwilligkeit der Lehrkräfte eine entscheidende Rolle, sodass auf eine Zufallsauswahl von Schulen und die häufig damit verbundenen Stichprobenausfälle auf Klassen- bzw. Schulebene verzichtet wurde. Innerhalb der Klassen wurden die Schülerinnen und Schüler per Zufall einer der drei Untersuchungsgruppen zugeteilt. Diese klasseninterne Randomisierung wird in Abschnitt 4 im Hinblick auf Leistungs- und Motivationsvariablen überprüft.

Fünf über einen mehrmonatigen Zeitraum verteilte Messzeitpunkte (Vortest, zwei formative Tests, Nachtest und Behaltenstest) führten allerdings zu krankheitsbedingten fehlenden Werten. Eine Analyse mit SPSS ergab, dass die Häufigkeit fehlender Werte nur bei zwei Variablen über 2 % liegt: Posttest (5.4 %) und Behaltenstest (8.8 %). Da diese beiden Variablen für die Hypothesenprüfung wichtig sind, könnte ein fallweiser Ausschluss zu einem Bias führen (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Vor Durchführung einer multiplen Imputation wurden die Fehlermuster unter Einbezug der Variablen Geschlecht, angestrebter Schulabschluss, Vornoten (Mathematik, Biologie, Deutsch), intrinsische Motivation (Interessenskomponente, Kompetenzkomponente), zu Hause gesprochene Sprache, Pretest, Posttest und Behaltenstest analysiert. Muster 1 (keine fehlenden Werte) hat eine Häufigkeit von ca. 86 % (vollständige Datensätze); Muster 7 (Behaltenstest fehlt) und 5 (Posttest fehlt) haben Häufigkeiten zwischen 4 und 7 %; alle anderen Muster haben eine Häufigkeit von unter 2 % und können vernachlässigt werden. In den beiden Mustern 7 und 5 finden sich keine systematischen Zusam-

menhänge der fehlenden Werte mit anderen Variablen, d. h. man kann mindestens MAR (missing at random) annehmen.

In SPSS wurde deshalb eine multiple Imputation vorgenommen. Folgende Variablen wurden als erklärende und zu imputierende Variablen definiert: Vornoten (Mathematik, Biologie, Deutsch), intrinsische Motivation (Interessenskomponente, Kompetenzkomponente), zu Hause gesprochene Sprache, Pretest, Posttest und Behaltenstest. Die Variablen Geschlecht und angestrebter Schulabschluss wurden nur als erklärende Variablen definiert, weil hier keine fehlenden Werte vorlagen. Für die Imputation wurde das automatische Verfahren gewählt (schließt eine Prüfung auf MAR ein). Es wurden insgesamt fünf imputierte Datensätze erzeugt. Die multivariaten Analysen (siehe Abschn. 4) wurden sowohl mit den Originaldaten als auch mit den fünf imputierten Datensätzen gerechnet. Es zeigten sich keine Unterschiede im Hinblick auf signifikante Effekte bei den einzelnen Variablen. Lediglich die Höhe der Beta-Werte variiert geringfügig.

4. Ergebnisse

4.1 Deskriptive Befunde

Tabelle 2 zeigt die deskriptiven Statistiken für alle erfassten Variablen in den nicht imputierten Originaldaten. Das Vorwissen der Schülerinnen und Schüler zu dieser Thematik ist erwartungsgemäß niedrig. Es findet insgesamt ein deutlicher Wissenszuwachs in der Unterrichtseinheit statt. Bei den Zeiten für Pre- und Posttest muss berücksichtigt werden, dass auch die motivationalen Variablen mit Moodle erfasst wurden und somit keine Zeitangaben für die reine Bearbeitung der Leistungstests vorliegen. Die Bearbeitungszeiten für die formativen Tests können dagegen als reine Bearbeitungszeit der Leistungstests interpretiert werden. Die Nutzung des ausführlichen Feedbacks liegt etwas über dem semantischen Median von 3 und streut relativ hoch.

In einem weiteren Schritt wurde analysiert, ob sich die drei Untersuchungsgruppen im Hinblick auf die Lernvoraussetzungen unterscheiden. Die einfaktorielle Varianzanalyse mit dem Faktor Versuchsgruppe führt zu keinem signifikanten Ergebnis (F-Werte werden nicht signifikant, $p > .05$), d. h. die Versuchsgruppen sind im Hinblick auf die erfassten Lernvoraussetzungen vergleichbar. Dies gilt auch für die nicht parametrischen Variablen Geschlecht und Migrationshintergrund (Kruskal-Wallis-Test unabhängiger Stichproben wird jeweils nicht signifikant). Ebenso wurde geprüft, ob sich die beiden Versuchsgruppen im Hinblick auf die Bearbeitungszeit der formativen Tests unterscheiden. Ein T-Test ergibt keinen signifikanten Mittelwertunterschied für die Bearbeitungszeit zwischen den beiden Rückmeldebedingungen (FLT 1: $T = 1.321$, $df = 163$, n. s.; FLT 2: $T = 1.106$, $df = 162$, n. s.). Der Levene-Test wird ebenfalls nicht signifikant, d. h. auch die Varianzen sind homogen (FLT 1: $F = 1.997$, n. s.; FLT 2: $F = 0.701$, n. s.). Man kann allerdings eine Tendenz hin zu längeren Bearbeitungszeiten und größerer Streuung bei der Versuchsbedingung mit ausführlicher Rückmeldung erkennen ($M_{Ausfl} = 21.00$

	N	Min	Max	M	SD
Intrinsische Motivation (Interesse)	256	1.00	5.00	3.10	0.83
Intrinsische Motivation (Kompetenz)	256	1.00	5.00	3.35	0.99
Mathematiknote	256	1.00	6.00	3.07	0.95
Biologie/PCB-Note	256	1.00	5.00	2.80	0.84
Deutschnote	256	1.00	5.00	2.93	0.75
Gesamtpunkte Pretest (max. 72)	256	11.00	51.00	27.95	6.49
Pretest Faktenwissen (max. 45)	256	2.00	30.00	17.33	4.84
Pretest Konzeptwissen (max. 27)	256	2.00	21.00	9.47	3.37
Zeit Pretest in Minuten	255	13.97	68.00	34.16	8.59
Gesamtpunkte FLT 1 (max. 46)	166	12.00	44.00	28.86	7.80
Zeit FLT 1 in Minuten	165	2.00	34.75	20.53	4.77
Gesamtpunkte FLT 2 (max. 52)	164	12.00	48.00	28.61	8.41
Zeit FLT 2 in Minuten	164	8.55	34.48	18.86	5.02
Nutzung des ausführlichen Feedbacks	79	1.00	5.00	3.35	0.95
Gesamtpunkte Posttest (max. 72)	247	16.00	67.00	43.53	10.77
Posttest Faktenwissen (max. 45)	247	10.00	43.00	28.19	7.39
Posttest Konzeptwissen (max. 27)	247	5.00	25.00	15.19	4.70
Zeit Posttest in Minuten	247	13.70	52.65	26.96	6.53
Gesamtpunkte Behaltenstest (max. 41)	242	0.00	32.00	13.13	7.46
Behaltenstest Faktenwissen (max. 17)	241	0.00	18.50	8.84	4.67
Behaltenstest Konzeptwissen (max. 24)	240	0.00	18.00	5.99	4.26

Tab. 2: Deskriptive Statistik

min vs. M_{Einf1} = 20.02 min; SD_{Ausf1} = 5.27 min vs. SD_{Einf1} = 4.16 min; M_{Ausf2} = 19.29 min vs. M_{Einf2} = 18.43 min; SD_{Ausf2} = 5.30 min vs. SD_{Einf1} = 4.71 min).

4.2 Effekte der Rückmeldevarianten

Um die Hypothesen zu prüfen, wurden multivariate Varianzanalysen gerechnet. Posttest und Behaltenstest dienten dabei als abhängige Variablen. Es wurden separate Analysen für die Effekte auf das Faktenwissen und das konzeptuelle Wissen gerechnet (Tab. 3). Feste Faktoren waren die Versuchsbedingung, die Abschlussart und das Geschlecht. Als Kovariaten wurden die Schulnoten und der Pretestwert (Faktenwissen bzw. konzeptuelles Wissen) eingesetzt.

		Faktenwissen			Konzeptuelles Wissen		
	Abhängige Variable	**F**	**p**	**Partielles Eta-Quadrat**	**F**	**p**	**Partielles Eta-Quadrat**
Korrigiertes Modell	Posttest	9.044	.000	.485	4.282	.000	.308
	Behaltenstest	11.702	.000	.549	7.310	.000	.432
Konstanter Term	Posttest	67.857	.000	.251	74.610	.000	.270
	Behaltenstest	34.550	.000	.146	45.588	.000	.184
Note Mathematik	Posttest	7.034	.009	.034	6.513	.011	.031
	Behaltenstest	4.178	.042	.020	5.078	.025	.025
Note Biologie	Posttest	8.202	.005	.039	.310	n. s.	
	Behaltenstest	14.425	.000	.067	11.150	.001	.052
Note Deutsch	Posttest	7.290	.008	.035	.059	n. s.	
	Behaltenstest	3.530	n. s.	.	.518	n. s.	
Pretest Faktenwissen bzw. konzept. Wissen	Posttest	37.191	.000	.155	16.784	.000	.077
	Behaltenstest	16.664	.000	.076	2.862	n. s.	
Abschluss	Posttest	18.082	.000	.152	14.928	.000	.129
	Behaltenstest	32.848	.000	.245	27.678	.000	.215
Geschlecht	Posttest	.074	n. s.		.689	n. s.	
	Behaltenstest	7.428	.007	.035	2.938	n. s.	
Treatment	Posttest	.107	n. s.		3.345	.037	.032
	Behaltenstest	1.002	n. s.		2.269	n. s.	
Abschluss * Geschl.	Posttest	2.130	n. s.		2.002	n. s.	
	Behaltenstest	6.350	.002	.059	2.933	n. s.	
Abschluss * Treatment	Posttest	.286	n. s.		2.259	n. s.	
	Behaltenstest	.192	n. s.		1.466	n. s.	
Geschlecht * Treatment	Posttest	.176	n. s.		.620	n. s.	
	Behaltenstest	.073	n. s.		1.373	n. s.	
Abschluss * Geschlecht * Treatment	Posttest	.769	n. s.		.424	n. s.	
	Behaltenstest	1.132	n. s.		1.926	n. s.	

Anmerkungen. Varianzaufklärung Faktenwissen Posttest R-Quadrat = .485 (korrigiert: .431); Varianzaufklärung Faktenwissen Behaltenstest: R-Quadrat = .549 (korrigiert: .502); Varianzaufklärung konzeptuelles Wissen Posttest: R-Quadrat = .308 (korrigiert: .236); Varianzaufklärung konzeptuelles Wissen Behaltenstest R-Quadrat = .432 (korrigiert: .373)

Tab. 3: Multivariate Varianzanalysen zur Prüfung der Treatmenteffekte (ausführliches Feedback vs. einfaches Feedback vs. Texte lesen) auf Post- und Behaltenstest (jeweils getrennte Analysen nach Faktenwissen und konzeptuellem Wissen)

Die höchsten Effekte sind bei der Abschlussart zu verzeichnen. Diese Variable wirkt sich erwartungsgemäß auf Posttest und Behaltenstest zum Faktenwissen und konzeptuellen Wissen aus. Bei den Schulnoten sind die Effekte geringer, vor allem die Mathematik- und Biologienote erklären einen kleinen Teil der Varianz. Der Pretest erklärt die Zunahme des Faktenwissens. Beim konzeptuellen Wissen wirkt sich der Pretest nur noch auf den Posttest aus. Zudem gibt es noch Haupt- und Interaktionseffekte des Geschlechts auf den Behaltenstest.

Ein signifikanter Einfluss der Versuchsgruppe findet sich lediglich beim konzeptuellen Wissen im Posttest. Der Effekt ist mit .03 zwar deutlich geringer als der Effekt des Pretests (der Effekt ist sowohl bei den Originaldaten als auch bei allen fünf imputierten Datensätzen signifikant und variiert zwischen .028 und .032). Er ist dennoch praktisch bedeutsam, wenn man bedenkt, dass es sich um die Variation eines Details innerhalb einer längeren Unterrichtseinheit handelt. Für die Prüfung der beiden Unterschiedshypothesen ist die Richtung des Effekts ausschlaggebend. Der paarweise Vergleich der geschätzten Randmittel ergibt, dass Lernende mit ausführlicher Rückmeldung (M = 13.8) signifikant schlechter abschneiden (p = .011) als Lernende mit einfacher Rückmeldung (M = 16.5). Es besteht jedoch kein Unterschied der beiden Versuchsbedingungen zur Kontrollgruppe (M = 14.5). Der Befund widerlegt somit beide Hypothesen.

4.3 Effekte der Rückmeldevarianten unter Berücksichtigung der Feedbacknutzung

Aus der Forschung zu Feedbackeffekten kann man Hinweise entnehmen, dass die Nutzung von Rückmeldungen den Effekt auf die Lernleistung moderiert. Die Angaben der Schülerinnen und Schüler zur Nutzung des Feedbacks werden deshalb herangezogen, um die Gruppe mit ausführlicher Rückmeldung entlang der z-standardisierten Werte post hoc in zwei Extremgruppen zu teilen: Nutzung der ausführlichen Rückmeldung vs. keine Nutzung der ausführlichen Rückmeldung.

Tabelle 4 zeigt die multivariate Varianzanalyse mit vier Versuchsgruppen für das Faktenwissen und das konzeptuelle Wissen. Abschlussart, Schulnoten und Pretest wurden wiederum als Kovariaten gewählt. Im Hinblick auf die Leistungsentwicklung beim Faktenwissen ergibt sich kein signifikanter Unterschied bei den vier post hoc gebildeten Versuchsgruppen. Dagegen sieht man wiederum einen Effekt der Versuchsbedingungen auf die Zunahme des Konzeptwissens, vor allem im Behaltenstest (die Tabelle zeigt die Werte der Originaldaten; in den fünf imputierten Datensätzen bestätigte sich der sign. Effekt des Behaltenstests auf dem 5 %-Signifikanzniveau; das Signifikanzniveau für den Effekt der Versuchsgruppen auf den Posttest schwankt zwischen .023 und .061). Angestrebter Schulabschluss, Schulnoten und Pretest erklären wiederum einen Teil.

Die geschätzten Randmittel sowie die paarweisen Vergleiche geben Auskunft über die Richtung des Effekts. Im Posttest unterscheidet sich die einfache Rückmeldung (M = 16.5) signifikant von der ausführlichen, jedoch nicht genutzten Rückmeldung (M = 13.0). Im Behaltenstest haben die ausführliche und gut genutzte Rückmeldung (M = 8.3) und

		Faktenwissen			Konzeptuelles Wissen		
	Abhängige Variable	**F**	**p**	**Partielles Eta-Quadrat**	**F**	**p**	**Partielles Eta-Quadrat**
Korrigiertes Modell	Posttest	7.160	.000	.497	3.437	.000	.321
	Behaltenstest	9.131	.000	.557	6.263	.000	.463
Konstanter Term	Posttest	66.508	.000	.253	63.682	.000	.245
	Behaltenstest	30.730	.000	.136	37.958	.000	.162
Note Mathematik	Posttest	7.471	.007	.037	5.573	.019	.028
	Behaltenstest	3.780	n. s.		4.294	.040	.021
Note Biologie	Posttest	8.990	.003	.044	.454	n. s.	
	Behaltenstest	13.735	.000	.065	12.007	.001	.058
Note Deutsch	Posttest	7.118	.008	.035	.349	n. s.	
	Behaltenstest	3.833	n. s.		1.561	n. s.	
Pretest Faktenwissen bzw. konzept. Wissen	Posttest	34.286	.000	.149	16.830	.000	.079
	Behaltenstest	17.145	.000	.080	3.789	n. s.	
Abschluss	Posttest	13.978	.000	.125	10.947	.000	.100
	Behaltenstest	28.190	.000	.223	15.744	.000	.138
Geschlecht	Posttest	.000	n. s.		.939	n. s.	
	Behaltenstest	4.370	.038	.022	2.043	n. s.	
Treatment (post hoc)	Posttest	.100	n. s.		2.722	.046	.040
	Behaltenstest	.690	n. s.		4.563	.004	.065
Abschluss * Geschl.	Posttest	1.863	n. s.		.864	n. s.	
	Behaltenstest	4.356	.014	.043	.357	n. s.	
Abschluss * Treatment	Posttest	.506	n. s.		1.696	n. s.	
	Behaltenstest	.498	n. s.		1.998	n. s.	
Geschlecht * Treatment	Posttest	.423	n. s.		.412	n. s.	
	Behaltenstest	.250	n. s.		2.052	n. s.	
Abschluss * Geschlecht * Treatment	Posttest	.630	n. s.		.461	n. s.	
	Behaltenstest	.598	n. s.		2.511	.023	.071

Anmerkungen. Varianzaufklärung Faktenwissen Posttest R-Quadrat = .485 (korrigiert: .431); Varianzaufklärung Faktenwissen Behaltenstest: R-Quadrat = .549 (korrigiert: .502); Varianzaufklärung konzeptuelles Wissen Posttest: R-Quadrat = .308 (korrigiert: .236); Varianzaufklärung konzeptuelles Wissen Behaltenstest R-Quadrat = .432 (korrigiert: .373)

Tab. 4: Multivariate Varianzanalysen zur Prüfung der Effekte post hoc gebildeter Treatmentgruppen (ausführliches Feedback genutzt vs. ausführliches Feedback nicht genutzt vs. einfaches Feedback vs. Texte lesen) auf Post- und Behaltenstest (jeweils getrennte Analysen nach Faktenwissen und konzeptuellem Wissen)

die einfache Rückmeldung (M = 7.5) jeweils eine signifikant höhere Punktzahl als das nicht genutzte ausführliche Feedback (M = 4.8) und die Kontrollgruppe (M = 5.5). D. h. unter Berücksichtigung der Feedbacknutzung bestätigt sich Hypothese 2, jedoch nicht Hypothese 1.

5. Diskussion

Die Beantwortung der Forschungsfrage führte zunächst zu einem erwartungswidrigen Befund. Die Versuchsgruppe mit einfacher Rückmeldung der Gesamtpunkte hatte einen signifikant höheren Wissenszuwachs als die Versuchsgruppe mit ausführlicher Rückmeldung pro Item bzw. als die Kontrollgruppe mit Lesetexten. In einer weiteren Analyse konnte gezeigt werden, dass dies vor allem mit der Nutzung der ausführlichen Rückmeldung durch die Schülerinnen und Schüler zusammenhängt. Wie lässt sich dieses Ergebnis in die bisherige Forschung zu Feedbackeffekten einordnen?

Der Befund bestätigt zunächst einmal Studien, die eine hohe Varianz der Nutzung von Rückmeldungen durch Lernende beschreiben (Timmers & Veldkamp, 2011; Rakoczy et al., 2013). Der Befund passt auch in das Bild der bisherigen Ergebnisse zu formativer Leistungsmessung in den naturwissenschaftlichen Fächern. Metaanalysen zeigen, dass geringere Effektstärken zu erwarten sind als in sprachlichen Fächern bzw. Mathematik (Kluger & DeNisi, 1996; Kingston & Nash, 2011). Dies hängt sowohl mit der Komplexität der naturwissenschaftlichen Wissensinhalte zusammen als auch mit der Tatsache, dass sich in Hauptfächern mit mehr Wochenstunden formative Lernverlaufsdiagnosen einfacher und effektiver umsetzen lassen. Unerwartet ist auch der Befund, dass die Leistungssteigerungen beim Faktenwissen nicht auf die formativen Tests zurückgeführt werden können. Eine mögliche Erklärung ist, dass das zu erlernende Faktenwissen sehr einfach war (z. B. Vogelschnäbel und Nahrung zuordnen, Flugarten benennen) und von allen Schülerinnen und Schülern in herkömmlicher Weise (z. B. Heft durchlesen, Fakten memorieren) vor dem benoteten Posttest gelernt bzw. wiederholt wurde.

Eine weitere Erklärung für den geringen Effekt der ausführlichen Rückmeldung ist die Feedbackinterventionstheorie (Kluger & DeNisi, 1996; Hattie & Timperley, 2007). Feedback auf der Ebene der Aufgabenbearbeitungsprozesse ist effektiver, weil sich bestimmte Aufgabenbearbeitungsstrategien auf viele weitere Aufgaben übertragen lassen. Im Projekt wurde sowohl für Aufgaben zu Faktenwissen als auch bei den Aufgaben zum konzeptuellen Wissen lediglich Feedback auf der Aufgabenlösungsebene realisiert (genaue Erläuterung der korrekten Lösung). Bei Feedback auf der Aufgabenlösungsebene führt vor allem sofortiges Feedback zu höheren Lernraten. Allerdings untersuchten wir einen sehr weiten Transfer auf den Posttest bzw. den Behaltenstest. D. h. ein stärker auf die Bearbeitungsprozesse bezogenes Feedback könnte eventuell zu höheren Behaltensleistungen führen.

Eine Schwierigkeit in der hier vorliegenden Studie ist mit Sicherheit der Umfang der ausführlichen Rückmeldungen. Bei der Durchführung der formativen Tests konnte im-

mer wieder beobachtet werden, dass die Schülerinnen und Schüler die Texte nicht oder nur kaum lesen und sofort weiterklicken. Dieses Problem der Umsetzung formativer Leistungsmessung wurde auch schon in vorangehenden Studien beobachtet (Yin et al., 2008). Eine weitere Erklärung hierfür ist die höhere kognitive Belastung durch ausführliche Rückmeldungen pro Item. Nach der Cognitive-Load-Theorie (Sweller, 1994) könnte die ausführliche Rückmeldung die nicht aufgabenbezogene kognitive Belastung erhöhen (extraneous load), wenn den Lernenden bereits durch eine einfache Rückmeldung der Korrektheit klar wurde, wo der Fehler lag. Für weiterführende Studien wären deshalb folgende Alternativen denkbar: Man könnte die Rückmeldungen kürzer und präziser formulieren. Ebenso könnte man auf eine Rückmeldung pro Item verzichten und am Ende des Tests in Abhängigkeit der Gesamtpunktzahl eine ausführliche Rückmeldung einblenden. Allerdings müsste man dann die formativen Moodle-Tests nach den zu diagnostizierenden Konzepten aufgliedern, d. h. ein formativer Test zu den Anpassungsmerkmalen, ein weiterer formativer Test zum Anpassungsbegriff usw. Auch der Umgang mit den Rückmeldungen müsste vor dem Einsatz der Tests mit den Schülerinnen und Schülern besprochen werden.

In dieser experimentellen Feldstudie wurde ein einzelner Aspekt formativer Leistungsdiagnostik zu einer bestimmten Lerndomäne gezielt variiert. Damit wurde auf Kritik an bisherigen Studien zu formativer Diagnostik reagiert (Bennett, 2011). Zieht man zudem in Betracht, dass die formative Leistungsmessung wiederum nur ein Element einer längeren Unterrichtseinheit war, konnte in dieser Studie ein deutlicher Effekt verschiedener Feedbackvarianten auf die Leistung im Behaltenstest bzw. Posttest gezeigt werden. Die Nutzung der formativen Rückmeldungen durch Lehrkräfte im nachfolgenden Unterricht wurde bewusst unterbunden, um keine weiteren Varianzquellen zu erzeugen. Für Folgestudien lässt sich damit die Hypothese formulieren, dass sowohl eine bessere Unterstützung der Lernenden bei der Nutzung des ausführlichen Feedbacks als auch eine Adaption des Unterrichts aufgrund des Feedbacks zu weitaus höheren Leistungseffekten führen könnten. Die formativen Testrückmeldungen könnten vor allem im Sinne der Forschung zum Konzeptwechsel in naturwissenschaftlichen Fächern (Duit & Treagust, 2010) für gezielte Gespräche herangezogen werden, um die noch vorhandenen Fehlvorstellungen der Lernenden zu hinterfragen bzw. ein korrektes Begriffsverständnis zu festigen. Ebenso könnte in Folgestudien untersucht werden, ob sich die Effekte des ausführlichen Feedbacks bei anderen Unterrichtsinhalten bzw. anderen Unterrichtsfächern ändern. Das Konzept der mehrschrittigen Testaufgaben sowie die Umsetzung in Moodle lassen sich auf alle Fächer mit komplexem Begriffswissen übertragen.

Neben der Hypothesenprüfung führte die Studie auch zu einer Reihe von praktischen Implikationen für die Umsetzung und Erforschung computergestützter, formativer Diagnostik. Zunächst einmal zeigte sich, dass Lernplattformen in Verbindung mit mobilen Endgeräten ein großes Potenzial für die Erweiterung der diagnostischen Möglichkeiten von Lehrkräften bieten und einen forschenden Blick auf die Lernentwicklung der eigenen Klasse eröffnen. Nach einer gewissen Einarbeitung bzw. nach dem Lösen erster technischer Schwierigkeiten waren die beteiligten Lehrkräfte an der digital dokumen-

tierten Lernentwicklung ihrer Schülerinnen und Schüler sehr interessiert. Die teilnehmenden Lehrkräfte würden elektronische Tests durchaus weiterhin einsetzen, wenn die entsprechende Ausstattung an ihren Schulen vorhanden wäre und sie damit umgehen könnten. Mit der zunehmenden Digitalisierung der Schulen eröffnen sich damit Handlungsoptionen für eine unterrichtsnahe, fachdidaktisch orientierte Forschung zu computergestützter, formativer Diagnostik.

Literatur

Anderson, K. T., Zuiker, S. J., Taasoobshirazi, G., & Hickey, D. T. (2007). Classroom Discourse as a Tool to Enhance Formative Assessment and Practise in Science. *International Journal of Science Education, 29*(14), 1721–1744.

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.

Baalmann, W., Frerichs, V., Weitzel, H., Gropengießer, H., & Kattmann, U. (2004). Schülervorstellungen zu Prozessen der Anpassung – Ergebnisse einer Interviewstudie im Rahmen der Didaktischen Rekonstruktion. *Zeitschrift für Didaktik der Naturwissenschaften, 10*, 7–28.

Bangert-Drowns, R. L., Kulik, C., Kulik, J. A., & Morgan, M. T. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research, 61*, 213–238.

Bennett, R. E. (2011). Formative Assessment: A critical review. *Assessment in Education, 18*(1), 5–25.

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education, 5*(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing the Theory of Formative Assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation* (4. Aufl.). Berlin: Springer.

Bürgermeister, A., Klimczak, M., Klieme, E., Rakoczy, K., Blum, W., Leiß, D., Harks, B., & Besser, M. (2011). Leistungsbeurteilung im Mathematikunterricht – Eine Darstellung des Projekts „Nutzung und Auswirkungen der Kompetenzmessung in mathematischen Lehr-Lernprozessen". *Schulpädagogik heute, 2*(3), 1–18.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The Development of a Two-Tier Multiple-Choice Diagnostic Instrument for Evaluating Secondary School Students' Ability to Describe and Explain Chemical Reactions Using Multiple Levels of Representation. *Chemistry Education Research and Practice, 8*(3), 293–307.

Chang, K.-E., Sung, Y.-T., Chang, R.-B., & Lin, S.-C. (2005). A New Assessment for Computer-Based Concept Mapping. *Educational Technology & Society, 8*(3), 138–148.

Clark, I. (2012). Formative Assessment: Assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205–249.

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research, 58*(4), 438–481.

Donovan, W. (2008). An Electronic Response System and Conceptests in General Chemistry Courses. *Journal of Computers in Mathematics and Science Teaching, 27*(4), 369–389.

Duit, R. (2003). Conceptual Change: A powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*, 671–688.

Duit, R., & Treagust, D. F. (2010). Conceptual Change: A powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*(6), 671–688.

Fraser, B. J., Walberg, H. J., Welch, W. W., & Hattie, J. A. C. (1987). Syntheses of Educational Productivity Research. *International Journal of Educational Research, 11*, 145–252.

Fuchs, L. S., & Fuchs, D. (1986). Effects of Systematic Formative Evaluation: A meta-analysis. *Exceptional Children, 53*, 199–208.

Furtak, E. M. (2012). Linking a Learning Progression for Natural Selection to Teachers' Enactment of Formative Assessment. *Journal of Research in Science Teaching, 49*(9), 1181–1210.

Halldén, O. (1988). The Evolution of the Species: Pupil perspectives and school perspectives. *International Journal of Science Education, 10*(5), 541–552.

Hattie, J. A. C. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement.* London/New York: Routledge, Taylor & Francis Group.

Hattie, J. A. C., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112.

Hussy, W., Schreier, M., & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften – Für Bachelor*. Berlin: Springer.

Jia, J., Chen, Y., Ding, Z., & Ruan, M. (2012). Effects of a Vocabulary Acquisition and Assessment System on Students' Performance in a Blended Learning Class for English Subject. *Computers & Education, 58*(1), 63–76.

Kingston, N., & Nash, B. (2011). Formative Assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Kluger, A. N., & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. *Psychological Bulletin, 119*(2), 254–284.

Lai, A.-F., & Chen, D.-J. (2010). Web-Based Two-Tier Diagnostic Test and Remedial Learning Experiment. *International Journal of Distance Education Technologies, 8*(1), 31–53.

Lin, S.-W. (2004). Development and Application of a Two-Tier Diagnostic Test for High School Students' Understanding of Flowering Plant Growth and Development. *International Journal of Science and Mathematics Education, 2*(2), 175–199.

Lipnevich, A. A., & Smith, J. K. (2009). Effects of Differential Feedback on Students' Examination Performance. *Journal of Experimental Psychology, 15*(4), 319–333.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau, 58*(2), 103–117.

Lysakowski, R. S., & Walberg, H. J. (1982). Instructional Effects of Cues, Participation, and Corrective Feedback: A quantitative synthesis. *American Educational Research Journal, 19*(4), 559–578.

Maier, U. (2010). Formative Assessment – Ein erfolgversprechendes Konzept zur Reform von Unterricht und Leistungsmessung? *Zeitschrift für Erziehungswissenschaft, 13*(2), 293–308.

Maier, U. (2014). Computergestützte, formative Leistungsdiagnostik in Primar- und Sekundarschulen. Ein Forschungsüberblick zu Entwicklung, Implementation und Effekten. *Unterrichtswissenschaft, 42*(1), 69–86.

Maier, U., Hofmann, F., & Zeitler, S. (2012). *Formative Leistungsdiagnostik – Grundlagen und Praxisbeispiele* (Schulmanagement-Handbuch 141). München: Oldenbourg.

McConnell, D. A., Steer, D. N., Owens, K. D., Knott, J. R., van Horn, S., Borowski, W., Dick, J., Foos, A., Malone, M., McGrew, H., Greer, L., & Heaney, P. J. (2006). Using Conceptests to Assess and Improve Student Conceptual Understanding in Introductory Geoscience Courses. *Journal of Geoscience Education, 54*(1), 61–68.

McKendree, J. (1990). Effective Feedback Content for Tutoring Complex Skills. *Human Computer Interaction, 5*, 381–413.

Mory, E. (1992). The Use of Informational Feedback in Instruction: Implications for future research. *Educational Training Research and Development, 40*(3), 5–20.

Nagata, N. (1993). Intelligent Computer Feedback for Second Language Instruction. *The Modern Language Journal, 77*(3), 330–339.

Nehm, R. H., & Reilly, L. (2007). Biology Majors' Knowledge and Misconceptions of Natural Selection. *BioScience, 57*(3), 263–272.

Posner, G., Strike, K., Hewson, P., & Gertzog, W. (1982). Accomodation of a Scientific Conception: Toward a theory of conceptual change. *Science Education, 66*, 211–227.

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written Feedback in Mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27*, 63–73.

Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The Interplay between Student Evaluation and Instruction – Grading and feedback in Mathematics classrooms. *Journal of Psychology, 216*(2), 111–124.

Randler, C. (2012). Field Experiments in Learning Research. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1293–1297). New York/Dordrecht/Heidelberg/London: Springer.

Randler, C., & Hummel, E. (2011). Vögel in unserer Umgebung? Wir erforschen die Geheimnisse des Fliegens. *RAAbits Realschule Biologie Tiere, 3*, 1–30.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and Issues in the Use of Concept Maps in Science Assessment. *Journal of Research in Science Teaching, 33*, 569–600.

Russell, M. K. (2010). Technology-Aided Formative Assessment of Learning: New developments and applications. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 125–138). New York/London: Routledge.

Steiner, G. (2001). Lernen und Wissenserwerb. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie – Ein Lehrbuch* (S. 139–205). Weinheim/Basel: Beltz.

Sweller, J. (1994). Cognitive Load Theory, Learning Difficulty and Instructional Design. *Learning and Instruction, 4*, 295–312.

Thissen-Roe, A., Hunt, E., & Minstrell, J. (2004). The DIAGNOSER Project: Combining assessment and learning. *Behavior Research Methods, Instruments, and Computers, 36*(2), 234–240.

Timmers, C., & Veldkamp, B. (2011). Attention Paid to Feedback Provided by a Computer-Based Assessment for Learning on Information Literacy. *Computers & Education, 56*(3), 923–930.

Treagust, D. F. (1988). Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science. *International Journal of Science Education, 10*(2), 159–169.

Wang, T.-H. (2011). Developing Web-Based Assessment Strategies for Facilitating Junior High School Students to Perform Self-Regulated Learning in an E-Learning Environment. *Computers & Education, 57*(2), 1801–1812.

Wild, K.-P., Krapp, A., Schiefele, U., Lewalter, D., & Schreyer, I. (1995). *Dokumentation und Analyse der Fragebogenverfahren und Tests. Berichte aus dem DFG-Projekt „Bedingungen und Auswirkungen berufsspezifischer Lernmotivation"*. Neubiberg: Universität der Bundeswehr München.

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., Tomita, M. K., & Young, D. B. (2008). On the Impact of Formative Assessment on Student Motivation, Achievement, and Conceptual Change. *Applied Measurement in Education, 21*(4), 335–359.

Zabel, J., & Gropengießer, H. (2010). Darwins konzeptuelle Landkarte: Lernfortschritt im Evolutionsunterricht. In U. Harms & I. Mackensen-Friedrichs (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik (Bd. 4), Heterogenität erfassen – individuell fördern im Biologieunterricht* (S. 209–224). Innsbruck: Studienverlag.

Abstract: In research on teaching and learning, formative assessments are considered to be an effective method of improving student performance. However, the effects vary strongly depending on learning content, diagnostic procedures, and the form of feedback. In a randomized experimental study conducted in biology classes, it was thus examined whether in the case of a computer-based, formative achievement test detailed in-depth feedback (treatment 1) would lead to better learning results than simple feedback (treatment 2). In a control group, students read the identical texts. 10 forms with a total of 261 students attending classes on the lower secondary level participated in the study. Results show that students receiving simple feedback achieve better results in both the post-test and the test of knowledge retained than students receiving treatment 1 or students of the control group. However, when looking more closely at the use made of the feedback, positive achievement effects can be shown for students getting in-depth feedback, too, compared to the control group.

Keywords: Formative Assessment, Computer, Feedback, Science Instruction, Student Performance

Anschrift der Autor_innen

Prof. Dr. Uwe Maier, Pädagogische Hochschule Schwäbisch Gmünd,
Institut für Erziehungswissenschaft/Empirische Schulforschung,
Oberbettringerstraße 200, 73525 Schwäbisch Gmünd, Deutschland
E-Mail: uwe.maier@ph-gmuend.de

Prof. Dr. Christoph Randler, Pädagogische Hochschule Heidelberg,
Institut für Naturwissenschaften/Didaktik der Biologie,
Im Neuenhainer Feld 561–2, 69198 Heidelberg, Deutschland
E-Mail: randler@ph-heidelberg.de

Dr. Nicole Wolf, Universität Erlangen-Nürnberg,
Lehrstuhl für Schulpädagogik,
Regensburger Straße 160, 90478 Nürnberg, Deutschland
E-Mail: nicole.wolf@fau.de